

# HEP-Help: a first-stop helpline for particle physics software

Jim Pivarski

Princeton University – IRIS-HEP

October 21, 2024

# HEP-Help: a first-stop helpline for particle physics software

Jim Pivarski

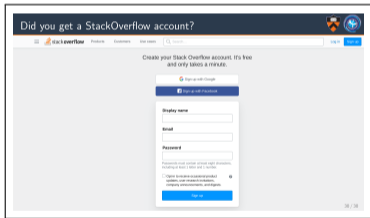
Princeton University – IRIS-HEP

October 21, 2024



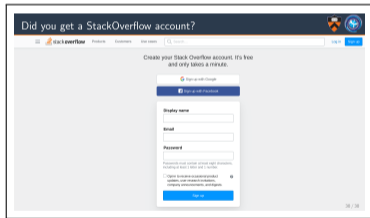


PyHEP 2019: I tried to get everybody to *use* StackOverflow, with tags to carve out our space within it.






PyHEP 2019: I tried to get everybody to *use* StackOverflow, with tags to carve out our space within it.



JLab Future Trends 2022: I acknowledged that it's not going well.

Addressing the disadvantages: user communication 

The problem is that we have too many ways to answer questions from users.

- ▶ GitHub issues and discussions: best so far, but distributed per-package
- ▶ Gitter: low-barrier chat, but also per-package
- ▶ Mattermost: CERN credentials are a barrier, but most LHC experiments are here
- ▶ Slack: required invitation is a barrier; mostly developers, anyway
- ▶ StackOverflow: good for cross-package discussions, but too diffuse in non-scientific world (when non-physicists answer questions, they're usually wrong)
- ▶ Scientific-Python.org Discourse and Discord: options under consideration

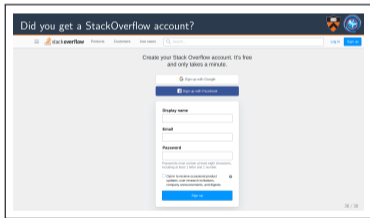
We would benefit by converging on one and sending users to that single forum.

27 / 33

# For a long time, I've wanted a "StackOverflow for HEP"



PyHEP 2019: I tried to get everybody to use StackOverflow, with tags to carve out our space within it.



JLab Future Trends 2022: I acknowledged that it's not going well.

Addressing the disadvantages: user communication

The problem is that we have too many ways to answer questions from users.

- ▶ GitHub issues and discussions: best so far, but distributed per-package
- ▶ Gitter: low-barrier chat, but also per-package
- ▶ Mattermost: CERN credentials are a barrier, but most LHC experiments are here
- ▶ Slack: required invitation is a barrier; mostly developers, anyway
- ▶ StackOverflow: good for cross-package discussions, but too diffuse in non-scientific world (when non-physicists answer questions, they're usually wrong)
- ▶ Scientific-Python.org Discourse and Discord: options under consideration

We would benefit by converging on one and sending users to that single forum.

27 / 33

Analysis Ecosystem II 2022 and PyHEP.dev 2023: Brainstorming sessions, never landed on a solution.

Analysis Ecosystem Workshop II

May 23 - 26, 2022  
UCLab  
Ameslab, Oak Ridge

Overview

Timeline

Contribution List

My Conference

My Contributions

Registration

Participant List

Code of Conduct

Travel

Accommodate

As part of the search for beyond the Standard Model physics, an array of next generation particle, nuclear and astroparticle experiments are under construction by global collaborations worldwide. These include the High-Luminosity Large Hadron Collider (HL-LHC) at CERN, the Deep Underground Neutrino Experiment (DUNE) at Fermilab, the Electron Ion Collider (EIC) at Brookhaven National Laboratory, the Facility for Antiproton and Ion Research at GSI, and many others.

PyHEP.dev 2023 "Python in HEP" Developer's Workshop

May 23 - 26, 2023  
UCLab, Oak Ridge

Overview

Timeline

Meeting Participants

Venue, Conference Fee, Hotel, Travel

Code of conduct

PyHEP.dev is an in-person, informal workshop for developers of Python software in HEP to plan a coherent roadmap and make priorities for the upcoming year. It complements the PyHEP Online workshop, which is intended for both developers and physicists.

Both PyHEP workshops are supported by the HEP Software Foundation (HSF). Further information is on



## Addressing the disadvantages: user communication



The problem is that we have *too many* ways to answer questions from users.

- ▶ [GitHub issues and discussions](#): best so far, but distributed per-package
- ▶ [Gitter](#): low-barrier chat, but also per-package
- ▶ [Mattermost](#): CERN credentials are a barrier, but most LHC experiments are here
- ▶ [Slack](#): required invitation is a barrier; mostly developers, anyway
- ▶ [StackOverflow](#): good for cross-package discussions, but too diffuse in non-scientific world (when non-physicists answer questions, they're usually wrong)
- ▶ [Scientific-Python.org](#) [Discourse](#) and [Discord](#): options under consideration

We would benefit by converging on one and sending users to that single forum.

# The ROOT Forum does not have this problem



The screenshot shows the ROOT Data Analysis Framework forum interface. At the top left is the ROOT logo with the text "Data Analysis Framework". To the right are search, menu, and user profile icons. Below the header is a navigation bar with "categories", "tags", "Latest" (underlined), "New (6)", "Unread (1)", "Hot", "Categories", "FAQ", and a "+ New Topic" button. The main content area has a table with columns for "Topic", "Replies", "Views", and "Activity".

Topic	Replies	Views	Activity
<p> <b>Newbie forum for when you're not sure</b></p> <p> News</p> <p>If you're new to ROOT, C++, data analysis etc, and you hesitate to ask your question, then please ask it in the Newbie section, where nice people help and we have special rules to be more welcoming. Don't hesitate, jus... read more</p>	2	17.9k	Sep 2018
<p> <b>Welcome to the ROOT forum!</b></p> <p> News</p> <p>ROOT is the tool of choice for analyzing, storing and graphing High Energy Physics data. Use this forum for getting help on ROOT or discussing its features, including math, cling, graphics, or the I/O. For documentatio... read more</p>	1	7.9k	Jun 2016

# The ROOT Forum does not have this problem



The screenshot shows the ROOT Data Analysis Framework forum interface. At the top, there is a search bar, a menu icon, and a user profile picture. Below this is a navigation bar with tabs for 'categories', 'tags', 'Latest', 'New (6)', 'Unread (1)', 'Hot', 'Categories', and 'FAQ'. A '+ New Topic' button is also present. The main content area displays a list of forum topics with columns for 'Topic', 'Replies', 'Views', and 'Activity'. Two topics are visible:

Topic	Replies	Views	Activity
<p><b>Newbie forum for when you're not sure</b></p> <p>News</p> <p>If you're new to ROOT, C++, data analysis etc, and you hesitate to ask your question, then please ask it in the Newbie section, where nice people help and we have special rules to be more welcoming. Don't hesitate, jus... read more</p>	2	17.9k	Sep 2018
<p><b>Welcome to the ROOT forum!</b></p> <p>News</p> <p>ROOT is the tool of choice for analyzing, storing and graphing High Energy Physics data. Use this forum for getting help on ROOT or discussing its features, including math, cling, graphics, or the I/O. For documentatio... read more</p>	1	7.9k	Jun 2016

- ▶ It's easy for newcomers to find, and ROOT team ensures that there's always someone "on shift" to answer questions.



# The ROOT Forum does not have this problem



The screenshot shows the ROOT Data Analysis Framework forum interface. At the top, there is a search bar, a menu icon, and a user profile picture. Below this, there are navigation tabs for 'categories', 'tags', 'Latest', 'New (6)', 'Unread (1)', 'Hot', 'Categories', and 'FAQ', along with a '+ New Topic' button. The main content area displays a list of forum topics with columns for 'Topic', 'Replies', 'Views', and 'Activity'.

Topic	Replies	Views	Activity
<p> <b>Newbie forum for when you're not sure</b></p> <p> News</p> <p>If you're new to ROOT, C++, data analysis etc, and you hesitate to ask your question, then please ask it in the Newbie section, where nice people help and we have special rules to be more welcoming. Don't hesitate, jus... read more</p>	2	17.9k	Sep 2018
<p> <b>Welcome to the ROOT forum!</b></p> <p> News</p> <p>ROOT is the tool of choice for analyzing, storing and graphing High Energy Physics data. Use this forum for getting help on ROOT or discussing its features, including math, cling, graphics, or the I/O. For documentatio... read more</p>	1	7.9k	Jun 2016

- ▶ It's easy for newcomers to find, and ROOT team ensures that there's always someone "on shift" to answer questions.
- ▶ Deep historical archive of past questions and answers.



Similarly, IRIS-HEP Slack, Coffea Users in CMS Mattermost, and some GitHub Discussions are very active. But the right forum can be hard to find, especially for problems that span multiple software packages.



Similarly, IRIS-HEP Slack, Coffea Users in CMS Mattermost, and some GitHub Discussions are very active. But the right forum can be hard to find, especially for problems that span multiple software packages.

Moving active communities is hard, and runs the risk of dispersing them instead.



Similarly, IRIS-HEP Slack, Coffea Users in CMS Mattermost, and some GitHub Discussions are very active. But the right forum can be hard to find, especially for problems that span multiple software packages.

Moving active communities is hard, and runs the risk of dispersing them instead.

**Better strategy:** make an entry point that

- ▶ shows people where a question has already been answered
- ▶ leads people to the right place to engage with already-active communities.

# New monkey wrench



● ChatGPT  
Search term

● StackOverflow  
Search term

+ Add comparison

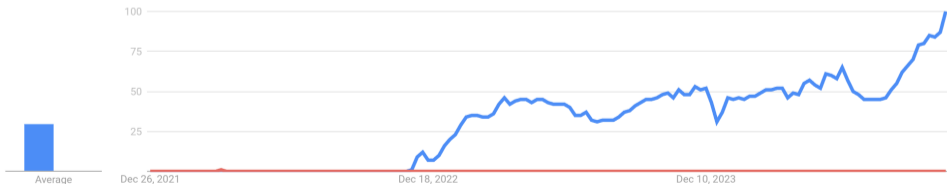
Worldwide ▾

1/1/22 - 10/16/24 ▾

All categories ▾

Web Search ▾

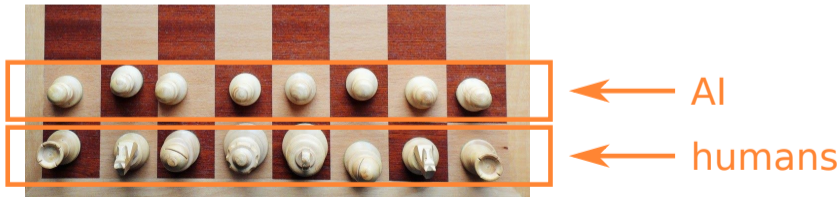
Interest over time ⓘ





Can a Large Language Model (LLM) be a first responder, either to answer questions or to send people to a forum where their question can be answered or has already been answered?

Can a Large Language Model (LLM) be a first responder, either to answer questions or to send people to a forum where their question can be answered or has already been answered?



# LLMs for HEP is a popular topic this year!



RAG/search	<a href="#">Leveraging Language Models to Navigate Conference Abstracts: An Open-Source Approach</a>	Gordon Watts	talk	next talk, here
RAG/search	<a href="#">AccGPT: A CERN Knowledge Retrieval Chatbot</a>	Florian Rehm, Juan Guijarro, Sofia Vallecorsa, Verena Kain	talk	20 minutes ago, rm 2A
RAG/search	<a href="#">Docu-Bot: AI assisted user support</a>	Jiri Chudoba	poster	maybe still up, Lobby
code review	<a href="#">Leveraging Language Models for Enhanced Code Review in Particle Physics Software Development</a>	Alexey Rybalchenko	poster	Tue 3pm, rm 4
domain-specific chat-bot	<a href="#">Xiwu: A basic flexible and learnable LLM for High Energy Physics</a>	Ke Li, Siyang Chen, Yiyu Zhang, Zhengde Zhang	poster	Tue 3pm, rm 4
domain-specific chat-bot	<a href="#">Boost physics study at HEP experiments with Dr. Sai</a>	<i>same authors</i> + Yipu Liao	poster	Tue 3pm, rm 4
general	<a href="#">Large Language Models in Physics</a>	Sarah Heim	plenary	Tue 11am <sub>8/21</sub>





## chATLAS in Production

- Since public beta launch ~2 weeks ago, there have been almost 1000 queries to the assistant
- Around 40% of these appear (from human inspection) to be decent answers
- The remainders are typically a refusal to answer due to lack of context, meaning that our retrieval system needs to be improved
- Already see improved context using Maximal Marginal Relevance retrieval
- Exploring:
  - More sophisticated automatic prompt generation
  - Passing query through ATLAS glossary
  - More physics-friendly embedding models

748	How can I download a file with Rucio ?	To download a file with Rucio, you can use the following command: "
749	Are you working?	The provided context does not contain specific information about wh
750	How can I store my analysis data at BNL ?	You can store your analysis data at BNL by utilizing the BNL-OSG2_GR
751	how can	I'm sorry, but the question seems to be incomplete. Please provide m
752	how can I add rooftop	I'm sorry, but the provided context from the ATLAS documentation an
753	how can I add a poi in roofit	I'm sorry, but the given context does not provide information related
754		What analysis was performed on the 2010 data in the ATLAS experim
755	How much data did we take in the year 2015?	Based on the available context, the document does not provide inform
756	What isolation working points are available for single lepton tri...	The available isolation working points for single lepton triggers includ
757	how to create an elementlink to an electron	To create an ElementLink to an electron, you can follow the example p
758	What is the Zprime particle in BSM theory?	The Zprime particle in BSM (Beyond the Standard Model) theory is a P
759	In a Tag and Probe trigger, how is the topo cut (e.g. 030RAB) us...	The information provided in the available context does not specificall
760	In a Tag and Probe trigger, what happens if the probe leg fails ...	If the probe leg fails all hypothesis steps in a Tag and Probe trigger. It
761	In a Tag and Probe trigger, what happens if the probe leg fails ...	In a Tag and Probe trigger scenario, if the probe leg fails all hypothesis
762		What analysis was performed on the 2010 data with 33pb <sup>-1</sup> , and wh
763	What are the main features of Atlas inner Detector ?	The main features of the ATLAS Inner Detector include: - improved ge
764	What are coherent prescale sets?	Coherent prescale sets are groups of triggers within the ATLAS experi
765	Describe coherent prescale sets	Coherent prescale sets, also referred to as prescaler groups, were intr
766	How long would it take to generate 1 million full simulation ev...	Based on the provided context, it is not possible to determine how lo
767	How long does the ATLAS fast chain event simulation take, per ...	Based on the provided context, the document mentions that the ATLAS
768	How long does a full delphes simulation of one full pileup even...	Based on the provided context, it is not specified how long a full Delph
769	How long does a full delphes simulation of one full pileup even...	Based on the available context, I cannot provide a specific answer to y
770	For a run of the ATLAS experiment, how can I check if a defect ...	To check if a defect assigned to a run for data quality purposes is t



## chATLAS in Production

- Since public beta launch ~2 weeks ago, there have been almost 1000 queries to the assistant
- Around 40% of these appear (from human inspection) to be decent answers
- The remainders are typically a refusal to answer due to lack of context, meaning that our retrieval system needs to be improved
- Already see improved context using Maximal Marginal Relevance retrieval
- Exploring:
  - More sophisticated automatic prompt generation
  - Passing query through ATLAS glossary
  - More physics-friendly embedding models

748	How can I download a file with Rucio ?	To download a file with Rucio, you can use the following command: "
749	Are you working?	The provided context does not contain specific information about wh
750	How can I store my analysis data at BNL ?	You can store your analysis data at BNL by utilizing the BNL-OSG2_GR
751	how can	I'm sorry, but the question seems to be incomplete. Please provide m
752	how can I add rooftop	I'm sorry, but the provided context from the ATLAS documentation an
753	how can I add a poi in roofit	I'm sorry, but the given context does not provide information related
754		What analysis was performed on the 2010 data in the ATLAS experim
755	How much data did we take in the year 2015?	Based on the available context, the document does not provide inform
756	What isolation working points are available for single lepton tri...	The available isolation working points for single lepton triggers includ
757	how to create an elementlink to an electron	To create an ElementLink to an electron, you can follow the example p
758	What is the Zprime particle in BSM theory?	The Zprime particle in BSM (Beyond the Standard Model) theory is a P
759	In a Tag and Probe trigger, how is the topo cut (e.g. 030RAB) us...	The information provided in the available context does not specificall
760	In a Tag and Probe trigger, what happens if the probe leg fails ...	If the probe leg fails all hypothesis steps in a Tag and Probe trigger. It
761	In a Tag and Probe trigger, what happens if the probe leg fails ...	In a Tag and Probe trigger scenario, if the probe leg fails all hypothesis
762		What analysis was performed on the 2010 data with 33pb <sup>-1</sup> , and wh
763	What are the main features of Atlas inner Detector ?	The main features of the ATLAS Inner Detector include: - improved ge
764	What are coherent prescale sets?	Coherent prescale sets are groups of triggers within the ATLAS experi
765	Describe coherent prescale sets	Coherent prescale sets, also referred to as prescaler groups, were intr
766	How long would it take to generate 1 million full simulation ev...	Based on the provided context, it is not possible to determine how lo
767	How long does the ATLAS fast chain event simulation take, per ...	Based on the provided context, the document mentions that the ATLAS
768	How long does a full delphes simulation of one full pileup even...	Based on the provided context, it is not specified how long a full Delph
769	How long does a full delphes simulation of one full pileup even...	Based on the available context, I cannot provide a specific answer to y
770	For a run of the ATLAS experiment, how can I check if a defect ...	To check if a defect assigned to a run for data quality purposes is t



## chATLAS in Production

- Since public beta launch ~2 weeks ago, there have been almost 1000 queries to the assistant
- Around 40% of these appear (from human inspection) to be decent answers
- The remainders are typically a refusal to answer due to lack of context, meaning that our retrieval system needs to be improved
- Already see improved context using Maximal Marginal Relevance retrieval
- Exploring:
  - More sophisticated automatic prompt generation
  - Passing query through ATLAS glossary
  - More physics-friendly embedding models

48	How can I download a file with Rucio ?	To download a file with Rucio, you can use the following command: "
49	Are you working?	The provided context does not contain specific information about wh
50	How can I store my analysis data at BNL ?	You can store your analysis data at BNL by utilizing the BNL-OSG2_GR
51	how can	
52	how can I add rooftop	
53	how can I add a poi in roofit	
54		
55	How much data did we take in the year 2015?	Based on the available context, the document does not provide inform
56	What isolation working points are available for single lepton tri...	The available isolation working points for single lepton triggers includ
57	how to create an elementlink to an electron	To create an ElementLink to an electron, you can follow the example p
58	What is the Zprime particle in BSM theory?	The Zprime particle in BSM (Beyond the Standard Model) theory is a P
59	In a Tag and Probe trigger, how is the topo cut (e.g. 030RAB) us...	The information provided in the available context does not specificall
60	In a Tag and Probe trigger, what happens if the probe leg fails ...	If the probe leg fails all hypothesis steps in a Tag and Probe trigger, it
61	In a Tag and Probe trigger, what happens if the probe leg fails ...	In a Tag and Probe trigger scenario, if the probe leg fails all hypothesi
62		What analysis was performed on the 2010 data with 33pb <sup>-1</sup> , and wh
63	What are the main features of Atlas inner Detector ?	The main features of the ATLAS Inner Detector include: - improved ge
64	What are coherent prescale sets?	Coherent prescale sets are groups of triggers within the ATLAS experi
65	Describe coherent prescale sets	Coherent prescale sets, also referred to as prescaler groups, were intr
66	How long would it take to generate 1 million full simulation ev...	Based on the provided context, it is not possible to determine how lon
67	How long does the ATLAS fast chain event simulation take, per ...	Based on the provided context, the document mentions that the ATLAS
68	How long does a full delphes simulation of one full pileup even...	Based on the provided context, it is not specified how long a full Delph
69	How long does a full delphes simulation of one full pileup even...	Based on the available context, I cannot provide a specific answer to y
70	For a run of the ATLAS experiment, how can I check if a defect ...	To check if a defect assigned to a run for data quality purposes is t

These are short questions;  
more like Google searches.



hep-help / hooray

Search: Type to search

Code Issues 1 Pull requests Discussions Actions Security Insights Settings

## Which versions of Awkward, Uproot, and Coffea align? #3

jpivarski started this conversation in General

**jpivarski** on Mar 6 Maintainer

I'm confused about Awkward 1 versus Awkward 2, Uproot 4 versus Uproot 5, and Coffea 0.7. Which versions of each package can be installed with each other?

1

1 comment · 1 reply

**hep-helper[bot]** on Mar 6

### Potentially useful sources

**21 September, 2020:** GitHub issue [scikit-hep/awkward#430](#). Installing (and then importing) awkward1 on Windows

Score: 90 out of 100

The conversation provides detailed information about the installation issues with Awkward 1 on Windows, including debugging steps and solutions. While it doesn't directly address the compatibility of Awkward 1, Awkward 2, Uproot 4, Uproot 5, and Coffea 0.7, it does offer insights into version dependencies and potential conflicts. Look for examples of troubleshooting steps, discussions on Windows installation, and the importance of compiling in Release mode to understand version compatibility and potential conflicts between different packages.

Category: General

Labels: None yet

1 participant

Notifications: Unsubscribe

You're receiving notifications because you're watching this repository.

Lock conversation

Transfer this discussion

Pin discussion

Pin discussion to General

Create issue from discussion

Delete discussion



A GitHub Actions bot, wired into GitHub Discussions/Issues. . .

- ▶ is free of charge and already has a nice UI (and CLI),
- ▶ handles authentication and most of us already have accounts,
- ▶ is not ephemeral/private: answered questions stay up for others to see,
- ▶ can be commented on by humans (e.g. "Careful! The above is wrong!"),
- ▶ symmetrically cross-reference any GitHub issues/PRs they link to,
- ▶ is implemented in GitHub Actions, which can run any code,
- ▶ which can securely access secrets, such as an OpenAI API key.



A GitHub Actions bot, wired into GitHub Discussions/Issues. . .

- ▶ is free of charge and already has a nice UI (and CLI),
- ▶ handles authentication and most of us already have accounts,
- ▶ is not ephemeral/private: answered questions stay up for others to see,
- ▶ can be commented on by humans (e.g. "Careful! The above is wrong!"),
- ▶ symmetrically cross-reference any GitHub issues/PRs they link to,
- ▶ is implemented in GitHub Actions, which can run any code,
- ▶ which can securely access secrets, such as an OpenAI API key.

The response can take up to a minute (much faster than a human responder).



A GitHub Actions bot, wired into GitHub Discussions/Issues. . .

- ▶ is free of charge and already has a nice UI (and CLI),
- ▶ handles authentication and most of us already have accounts,
- ▶ is not ephemeral/private: answered questions stay up for others to see,
- ▶ can be commented on by humans (e.g. “Careful! The above is wrong!”),
- ▶ symmetrically cross-reference any GitHub issues/PRs they link to,
- ▶ is implemented in GitHub Actions, which can run any code,
- ▶ which can securely access secrets, such as an OpenAI API key.

The response can take up to a minute (much faster than a human responder).

**Discussions or Issues?** Discussions are threaded (3 levels) with up/down votes, but Issues can ask users to fill out a structured form.

# How it works: using a CI workflow to respond to user posts



```
name: answer-query
on:
  discussion:
    types: [created, edited]

jobs:
  answer-query:
    name: answer-query
    runs-on: ubuntu-latest
    steps:
      - name: Git checkout
        uses: actions/checkout@v4
        with:
          fetch-depth: 0
      - name: Get Python
        uses: actions/setup-python@v5
        with:
          python-version: "3.11"
      - name: Install dependencies
        run: |
          python -m pip install \
            -r requirements.txt
      - name: Get vector store
        shell: bash
        run: |
          export TAG=`git describe --abbrev=0 --tags`
          wget https://github.com/hep-help/hooray/ \
            releases/download/$TAG/hep-help-db.zip
          unzip hep-help-db.zip
      - name: Produce response
        shell: bash
        env:
          OPENAI_API_KEY: ${{ secrets.OPENAI_API_KEY }}
          BODY: ${{ github.event.discussion.body }}
        run: |
          echo "$BODY" | python answer-query.py > ./text.md
      - name: Post response
        shell: bash
        env:
          APP_PRIVATE_KEY: ${{ secrets.APP_PRIVATE_KEY }}
          DISCUSSION_ID: ${{ github.event.discussion.node_id }}
        run: |
          echo "$APP_PRIVATE_KEY" > ./key.pem
          python comment-on-discussion.py
```





- ▶ Gather documents from many sources: GitHub/GitLab, Slack, (public?) Mattermost, Gitter, StackOverflow, Discord, documentation websites, HSF-Training tutorial materials. . .
- ▶ Understand how LLM technology works to improve responses
- ▶ Streamline the user interface
- ▶ Advertise widely

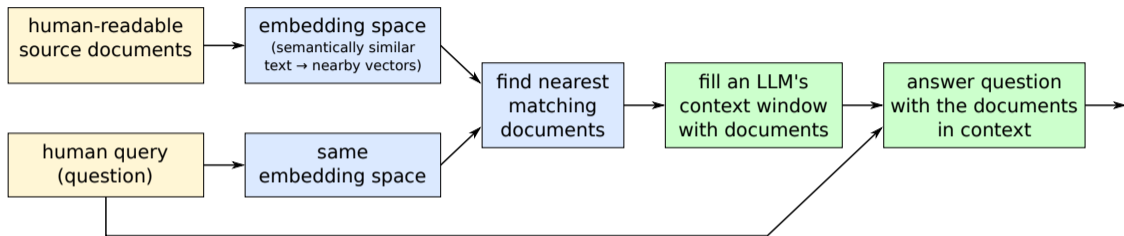


- ▶ Gather documents from many sources: GitHub/GitLab, Slack, (public?) Mattermost, Gitter, StackOverflow, Discord, documentation websites, HSF-Training tutorial materials. . .

- ▶ Understand how LLM technology works to improve responses

(the rest of this talk)

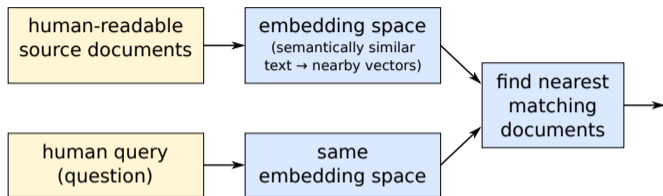
- ▶ Streamline the user interface
- ▶ Advertise widely



## Retrieval-Augmented Generation (RAG):

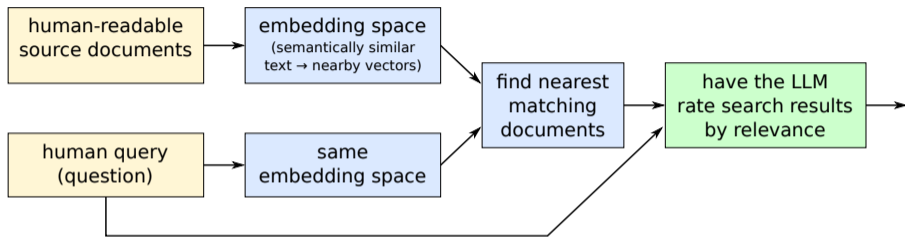
An LLM is better able to answer questions if it has relevant information in its (limited-size) context window.

Get information by passing documents and query through the same neural network; in that embedding space, similar vectors are semantically similar text.



## Just semantic search:

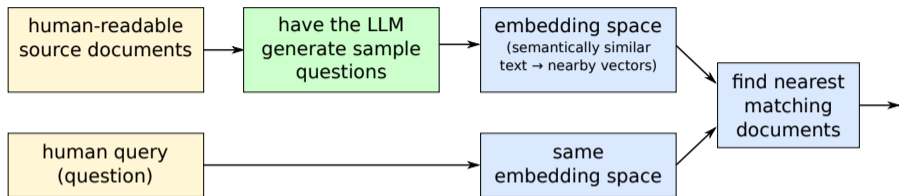
Maybe we don't need the LLM at all! If we can match a query to semantically similar documents, perhaps we should just recommend these documents.



## Semantic search with re-ranking:

Some embedding space matches aren't actually related to the query, even though they touch on the same concepts.

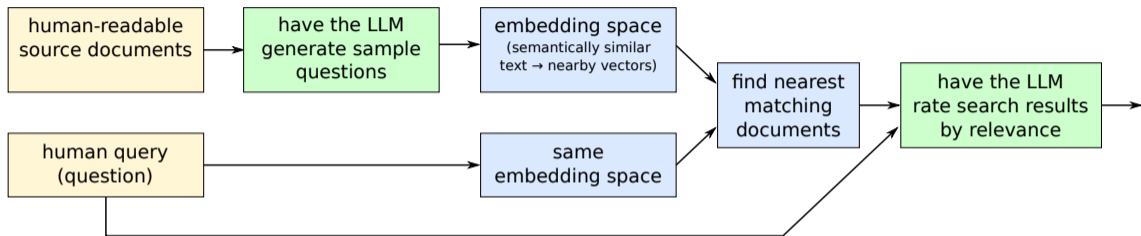
Use the LLM in a limited way: have it rate and describe how relevant each match is to the human query.



## Better targets in embedding space:

The human query is a question, and a question is a better semantic match to a question than a document that would answer that question.

Use the LLM in a limited way: have it generate possible questions about the source documents.



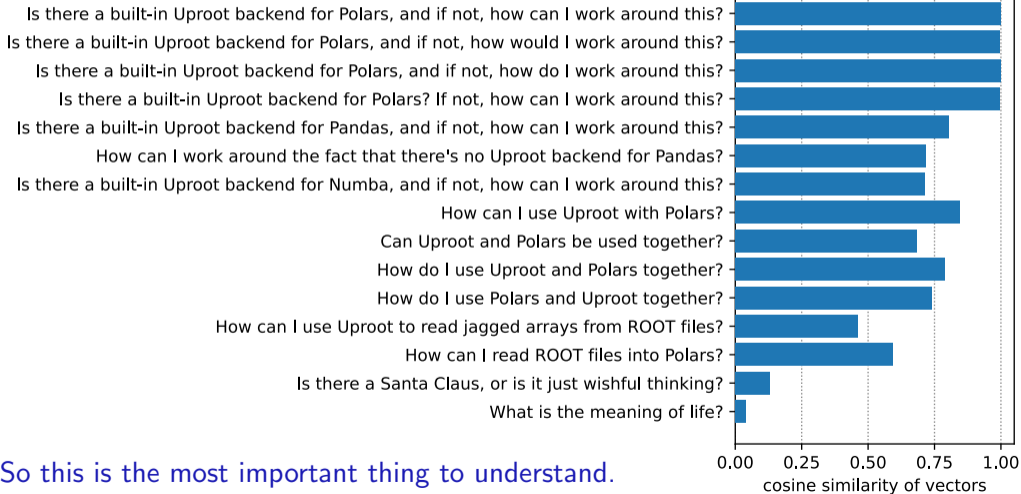
Combine #3 and #4:

Note: the LLM cost for #3 scales with the number of documents, but #4 scales with the number of human queries.

# Without a good embedding, we won't find the right documents



text-embedding-3-large (3072 dim)



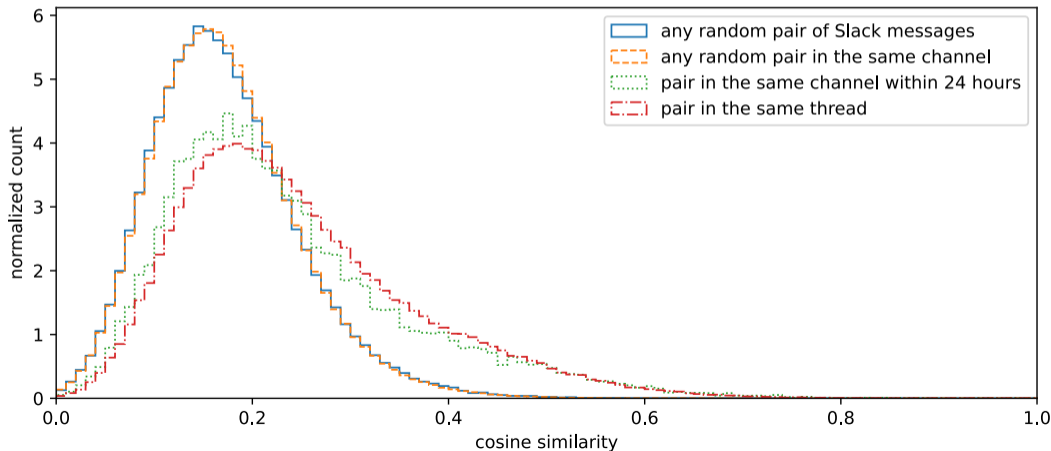
So this is the most important thing to understand.



# How much wiggle room is there between signal and background?



Consecutive Slack messages (IRIS-HEP Slack) are a large sample of question-answer pairs: the “signal.” Non-consecutive ones are “background.”



# Cleaner set of question-answer pairs: StackOverflow



stackoverflow Products OverflowAI  5,859 3 36 50 +1340

Home Questions Staging Ground Tags Saves Users Companies LABS Jobs Discussions COLLECTIVES TEAMS overflow AI

Now available on Stack Overflow for Teams! AI features where you work, search, IDE, and chat. Learn more Explore Teams Looking for your Teams?

## All Questions [Ask Question](#)

Tagged with [awkward-array](#) or [uproot](#)

156 questions Newest Active Bountied Unanswered More [Filter](#)

1 vote [1 answer](#)  
28 views  
`python` `uproot`  
[browndragon 55](#) asked Aug 19 at 17:40

1 vote [2 answers](#)  
165 views  
`python` `awkward-array`  
[Matt Bellis 340](#) asked Jun 22 at 23:02

1 vote 0 answers  
57 views  
`numpy` `numpy-ndarray` `array-broadcasting` `awkward-array`  
[Nico 11](#) asked Jun 17 at 14:49

1 vote [1 answer](#)  
44 views  
`pandas` `group-by` `awkward-array`  
[Victor Ruelas 13](#) asked Apr 22 at 11:06

2 votes [2 answers](#)  
192 views  
`pytorch` `awkward-array`  
[Teddy Curtis 21](#) asked Mar 22 at 11:11

### The Overflow Blog

- What launching rockets taught this CTO about hardware observability
- The team behind Unity 6 explains the new features aimed at helping developers

### Featured on Meta

- Preventing unauthorized automated access to the network
- Upcoming initiatives on Stack Overflow and across the Stack Exchange network...
- Feedback Requested: How do you use the tagged questions page?
- Proposed designs to update the homepage for logged-in users

### Hot Meta Posts

- 58 Should I not flag users with very questionable SG reviews?
- 17 Curses, [cursor]! Should we burninate and break the spell?

### Custom Filters

[Create a custom filter](#)

### Watched Tags

`awkward-array` `uproot` [edit](#)

# Cleaner set of question-answer pairs: StackOverflow



The screenshot shows the Stack Overflow interface. At the top, there's a search bar with the text "[awkward-array] or [uproot]". Below the search bar, the page title is "All Questions" with an "Ask Question" button. The page is tagged with "awkward-array" and "uproot". There are 156 questions listed, with filters for "Newest", "Active", "Bountied", "Unanswered", and "More".

The first question is "rebinning hist object from linear bins to log(x) bins in scikit-learn hist" with 1 vote and 1 answer. The second is "Accessing elements of an awkward array that are not a passed-in index" with 2 answers. The third is "Awkward Array broadcasting and linear indexing, reshaping an Awkward Array" with 0 answers. The fourth is "Slicing Multindex pandas Dataframe" with 1 answer. The fifth is "Converting from awkward arrays into torch arrays" with 2 answers.

On the right side, there's a "The Overflow Blog" section with two articles: "What launching rockets taught this CTO about hardware observability" and "The team behind Unity 6 explains the new features aimed at helping developers". Below that is "Featured on Meta" with three items: "Preventing unauthorized automated access to the network", "Upcoming initiatives on Stack Overflow and across the Stack Exchange network...", and "Feedback Requested: How do you use the tagged questions page?". There's also a "Hot Meta Posts" section with two posts: "Should I not flag users with very questionable SG reviews?" and "Curses, [cursor]! Should we burninate and break the spell?".

At the bottom right, there's a "Custom Filters" section with a "Create a custom filter" button, and a "Watched Tags" section with "awkward-array" and "uproot" tags.

Eliminating out-of-date answers, we have 85 question-answer pairs.

# Cleaner set of question-answer pairs: StackOverflow



The screenshot shows the Stack Overflow interface. At the top, there's a search bar with the query "[awkward-array] or [uproot]". The main content area is titled "All Questions" and shows a list of 156 questions. The questions are sorted by "Newest".

Questions listed:

- 1 vote**, 1 answer, 28 views: [rebinning hist object from linear bins to log\(x\) bins in scikit-learn hist](#). Asked by [browndragon](#) 55 on Aug 19 at 17:40.
- 1 vote**, 2 answers, 165 views: [Accessing elements of an awkward array that are not a passed-in index](#). Asked by [Matt Bellis](#) 340 on Jun 22 at 23:02.
- 1 vote**, 0 answers, 57 views: [Awkward Array broadcasting and linear indexing, reshaping an Awkward Array](#). Asked by [Nico](#) 11 on Jun 17 at 14:49.
- 1 vote**, 1 answer, 44 views: [Slicing MultiIndex pandas Dataframe](#). Asked by [Victor Ruelas](#) 13 on Apr 22 at 11:06.
- 2 votes**, 2 answers, 192 views: [Converting from awkward arrays into torch arrays](#). Asked by [Teddy Curtis](#) 21 on Mar 22 at 11:11.

On the right side, there are sections for "The Overflow Blog", "Featured on Meta", and "Hot Meta Posts".

Eliminating out-of-date answers, we have 85 question-answer pairs.

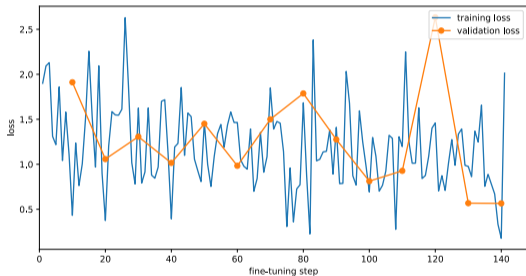
(Almost all the answers were written by me.)

# Could we even use them to *fine-tune* ChatGPT?



50%, 25%, 25% training, validation, test

Loss versus training step is not promising:

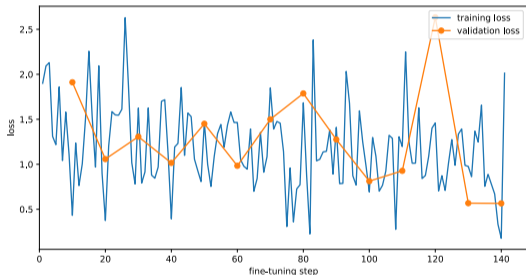


# Could we even use them to *fine-tune* ChatGPT?



50%, 25%, 25% training, validation, test

Loss versus training step is not promising:



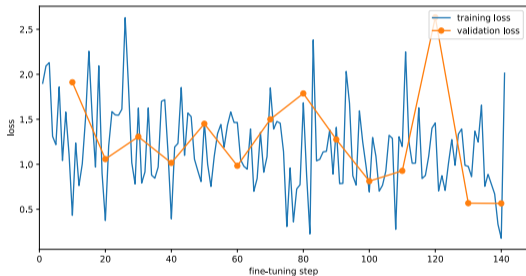
But whereas untuned ChatGPT always attempts non-vectorized answers to user questions, fine-tuned always attempts vectorized (though usually wrong) ones.

# Could we even use them to *fine-tune* ChatGPT?



50%, 25%, 25% training, validation, test

Loss versus training step is not promising:



But whereas untuned ChatGPT always attempts non-vectorized answers to user questions, fine-tuned always attempts vectorized (though usually wrong) ones.

Untuned ChatGPT:

```
# Loop over each sublist and replace the elements
for i, (sublist, mod, mod_idx) in enumerate(zip(
    A_final, A_mod, A_mod_idx
)):
    for idx, value in zip(mod_idx, mod):
        sublist[idx] = value

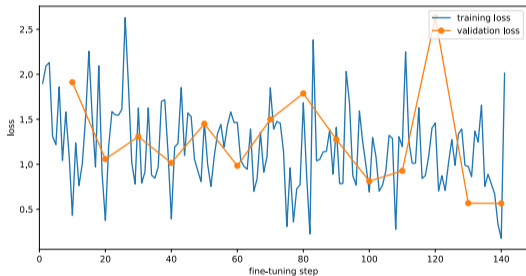
# Convert the modified list back to an Awkward Array
A_final = ak.Array(A_final)
```

# Could we even use them to *fine-tune* ChatGPT?



50%, 25%, 25% training, validation, test

Loss versus training step is not promising:



But whereas untuned ChatGPT always attempts non-vectorized answers to user questions, fine-tuned always attempts vectorized (though usually wrong) ones.

## Untuned ChatGPT:

```
# Loop over each sublist and replace the elements
for i, (sublist, mod, mod_idx) in enumerate(zip(
    A_final, A_mod, A_mod_idx
)):
    for idx, value in zip(mod_idx, mod):
        sublist[idx] = value

# Convert the modified list back to an Awkward Array
A_final = ak.Array(A_final)
```

## Fine-tuned ChatGPT: (ak.with\_item does not exist)

```
>>> import awkward as ak
>>> A = ak.Array([[21.1, 22.1, 43.1], [73.1, 30.4], [43.1]])
>>> A_mod_idx = ak.Array([[0, 2], [1], []])
>>> mask = ak.is_none(ak.values_astype(
...     ak.local_index(A, axis=1)[: , :, None] == A_mod_idx,
...     bool,
... ), axis=2)
>>> mask
<Array [[False, True, ...], ..., [True]] type='3 * var * bool'>

# Now we can use ak.with_item to put the new values into the
# places where the mask is False:

>>> ak.with_item(A, A_mod, mask)
<Array [[1.1, 22.1, 2.1], ..., [43.1]] type='3 * var * float64'>
```

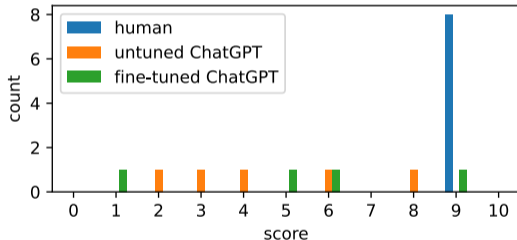




I made a tool to conceal the author of an answer and “graded” answers on a scale from 0 to 10. (I don't *consciously* remember writing these answers.)

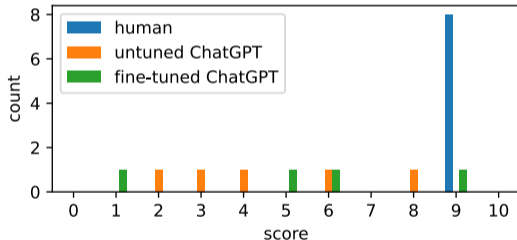


I made a tool to conceal the author of an answer and “graded” answers on a scale from 0 to 10. (I don't *consciously* remember writing these answers.)





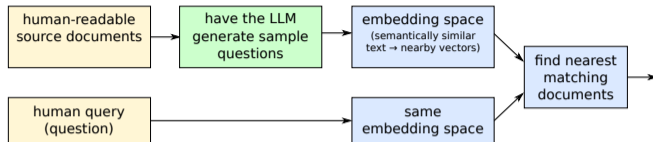
I made a tool to conceal the author of an answer and “graded” answers on a scale from 0 to 10. (I don’t *consciously* remember writing these answers.)



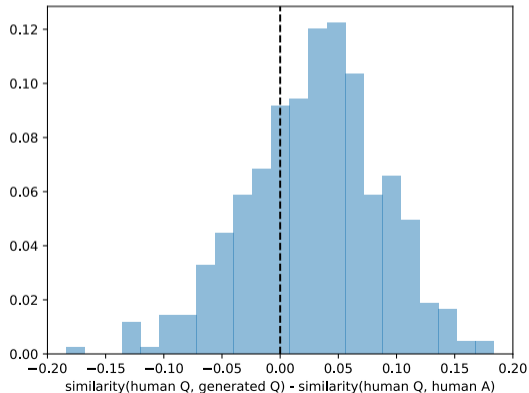
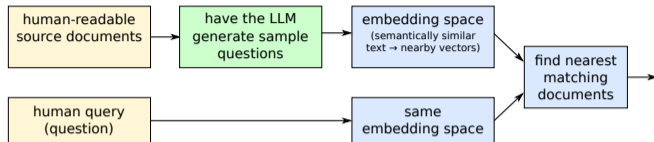
It takes 5–10 minutes to “grade” an answer; this can’t be done in bulk.



# What about using the LLM as a question-generator?

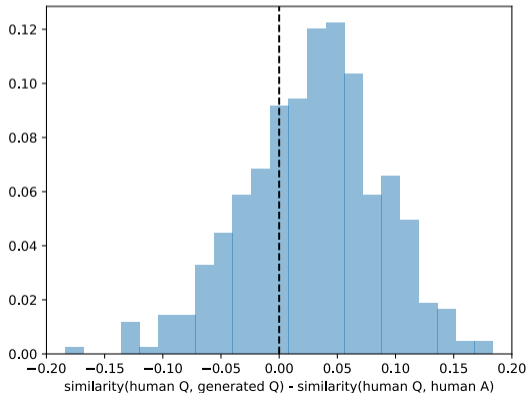
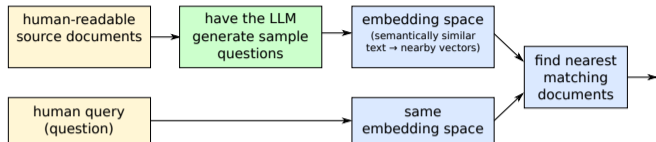


# What about using the LLM as a question-generator?



The similarity of a human query and an LLM-generated question is slightly better than a human query and the human source documents (classic RAG).

# What about using the LLM as a question-generator?



The similarity of a human query and an LLM-generated question is slightly better than a human query and the human source documents (classic RAG).

(Is it enough to matter?)



- ▶ We do need a “funnel” to guide physicists toward help with software, especially for questions that cross package boundaries.



- ▶ We do need a “funnel” to guide physicists toward help with software, especially for questions that cross package boundaries.
- ▶ A GitHub Actions bot is a good interface/technology for that.





- ▶ We do need a “funnel” to guide physicists toward help with software, especially for questions that cross package boundaries.
- ▶ A GitHub Actions bot is a good interface/technology for that.
- ▶ In principle, LLMs can help, too!
  - ▶ It's a popular topic.
  - ▶ But how this will work is unclear.



- ▶ We do need a “funnel” to guide physicists toward help with software, especially for questions that cross package boundaries.
- ▶ A GitHub Actions bot is a good interface/technology for that.
- ▶ In principle, LLMs can help, too!
  - ▶ It's a popular topic.
  - ▶ But how this will work is unclear.
- ▶ It's essentially a search problem, so understanding and optimizing similarity in the embedding space is crucial.